Image Caption Generation with Recurrent Neural Networks

Siddartha Ravichandran , Sonam Choudhary

THE UNIVERSITY OF UTAH

OVERVIEW

We aim to recognize image features and generate accurate, syntactically reasonable text descriptions for that images. Most importantly our goal is to predict a corresponding caption or at the least score the image containing common objects in their natural context against 5 possible captions with the hope that the closest caption (or the one that actually fits the image) is scored highest.

DATASET

- Microsoft Common Object in Context (MS COCO) dataset
  - Total 123278 images
  - each image labelled with 5 captions
  - Around 80k for training
  - Around 40K for testing
- Designed for the detection and segmentation of objects occurring in their natural context.
- MS COCO has fewer categories than ImageNet and SUN, it has more instances per category.
- In comparison to other datasets MS COCO has both more categories and instances.

- **EXAMPLE:**

Captions:

1. A tennis player in front of other people on the side lines.
2. A tennis player is wiping his racket off with a towel.
3. A man in all white holding a racket.
4. A good looking tennis player and his friends.
5. man holding a towel and a tennis racket, men in the back looking through backpacks and sitting on chairs.

TECHNOLOGY

- Neural Network Package: **TORCH**
- RNN package : Element-Research/rnn
- Run on 8-core system, with Tesla K80 GPU
- Loadcaffe: VGG CNN model

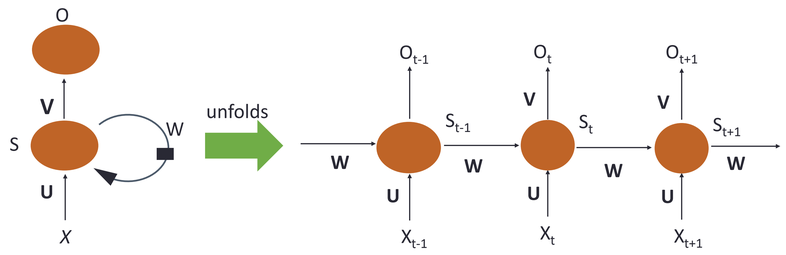
MODEL

Our alignment model is based on a combination of Convolutional Neural Networks over image regions and Recurrent Neural Networks over sentences. We are given an image and its corresponding sequence during the training stage. So essentially we want to learn the sequence of the caption given the context of the image. Assuming we can map our feature representation of images and the individual words of the caption to a common space, we can feed into the RNN a sequence in the following order:

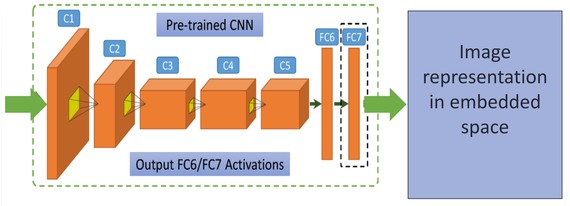
[Image] → [Caption_0] → [Caption_1] → [Caption_2] . . . → [Caption_n]

where [ ] represents a common embedded space.

RNN Unit:

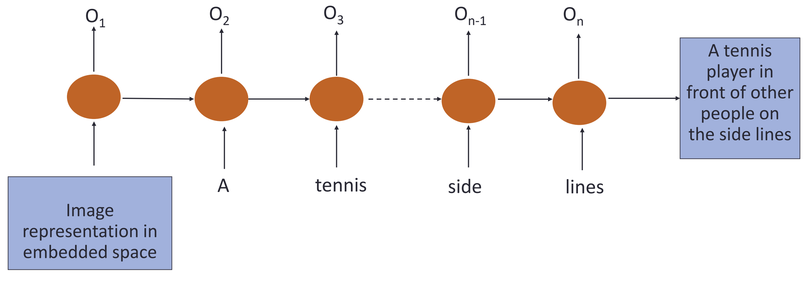


Common Step:

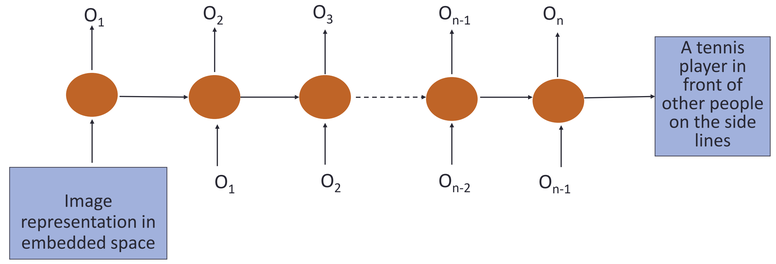


Convolution Neural Network

Training Step:



Prediction Step:



RESULTS

Caption	Score
Man riding a motor bike on a dirt road on the countryside.	6.12
Man is siting on a bike.	5.98
A man is walking on the road.	11.34
A boy is playing with a ball.	3.58
A man is sitting on a bench.	7.54

Caption	Score
A person holding a cell phone in their hand.	8.09
A man in wearing a black shirt.	9.82
A person is talking on phone.	14.28
A man is eating food.	10.36
A book is kept on the table.	4.65

FURTURE WORK

- An RNN conditions the generation of each word on the entire previous history of words generated. A Markov chain only conditions on a fixed window. Perhaps a particular RNN will learn to truncate its conditioning context and behave as a Markov chain, or perhaps not; but RNNs in general certainly can generate languages that Markov chains cannot.
- Studying and applying other RNN concepts like LSTMs, Gated Recurrent Units.